JB
Journals.ASM.org

Comparative Genomic Analyses of 17 Clinical Isolates of *Gardnerella vaginalis* Provide Evidence of Multiple Genetically Isolated Clades Consistent with Subspeciation into Genovars

Azad Ahmed,[a] Josh Earl,[a] Adam Retchless,[a]* Sharon L. Hillier,[b,c] Lorna K. Rabe,[b] Thomas L. Cherpes,[b] Evan Powell,[a] Benjamin Janto,[a] Rory Eutsey,[a] N. Luisa Hiller,[a] Robert Boissy,[a]* Margaret E. Dahlgren,[a] Barry G. Hall,[a,d] J. William Costerton,[a,e,f] J. Christopher Post,[a,f,g] Fen Z. Hu,[a,f,g] and Garth D. Ehrlich[a,f,g]

Center for Genomic Sciences, Allegheny Singer Research Institute, Pittsburgh, Pennsylvania, USA[a]; Magee-Womens Research Institute, Pittsburgh, Pennsylvania, USA[b]; University of Pittsburgh, Pittsburgh, Pennsylvania, USA[c]; Bellingham Institute, Bellingham, Washington, USA[d]; Department of Orthopaedics, Allegheny General Hospital, Pittsburgh, Pennsylvania, USA[e]; and Departments of Microbiology and Immunology[f] and Otolaryngology and Head and Neck Surgery,[g] Drexel University College of Medicine, Allegheny Campus, Pittsburgh, Pennsylvania, USA

***Gardnerella vaginalis* is associated with a spectrum of clinical conditions, suggesting high degrees of genetic heterogeneity among stains. Seventeen *G. vaginalis* isolates were subjected to a battery of comparative genomic analyses to determine their level of relatedness. For each measure, the degree of difference among the *G. vaginalis* strains was the highest observed among 23 pathogenic bacterial species for which at least eight genomes are available. Genome sizes ranged from 1.491 to 1.716 Mb; GC contents ranged from 41.18% to 43.40%; and the core genome, consisting of only 746 genes, makes up only 51.6% of each strain's genome on average and accounts for only 27% of the species supragenome. Neighbor-grouping analyses, using both distributed gene possession data and core gene allelic data, each identified two major sets of strains, each of which is composed of two groups. Each of the four groups has its own characteristic genome size, GC ratio, and greatly expanded core gene content, making the genomic diversity of each group within the range for other bacterial species. To test whether these 4 groups corresponded to genetically isolated clades, we inferred the phylogeny of each distributed gene that was present in at least two strains and absent in at least two strains; this analysis identified frequent homologous recombination within groups but not between groups or sets. *G. vaginalis* appears to include four nonrecombining groups/clades of organisms with distinct gene pools and genomic properties, which may confer distinct ecological properties. Consequently, it may be appropriate to treat these four groups as separate species.**

Gardnerella vaginalis is a facultative anaerobic coccobacillus that has a Gram-positive cell wall structure (28) but stains as a Gram-indeterminate bacterium because its cell wall is very thin, which allows it to appear as either Gram positive or Gram negative. *G. vaginalis* has been recovered from women with upper reproductive tract infections, including endometritis and pelvic inflammatory disease, as well as asymptomatic women, but it is most often observed and isolated as the dominant microorganism in the vaginal microflora of women suffering from bacterial vaginosis (BV), a highly prevalent disease affecting 10% to 40% of women of reproductive age (2, 13, 42, 68). BV is characterized by a malodorous vaginal discharge (65) as well as reduced vaginal acidity and the presence of clue cells (1). Clue cells are bacterium-covered human epithelial cells present in the vaginal discharge of women with BV. A Gram-stained BV sample usually shows a high prevalence of *Gardnerella* morphotypes and a deficiency of the lactobacillus morphotypes that are normally found in the vagina; these changes form the basis of the Nugent score for diagnosis of BV (33). BV is associated with increased risks for preterm delivery (33, 53), intrauterine growth retardation (22), pelvic inflammatory disease (26), postpartum endometritis (76), and HIV infection (69). Women with laboratory evidence of BV, but no symptoms, may still be at increased risk for adverse health outcomes. In addition, *G. vaginalis* has been associated with extrareproductive tract infections, including vertebral osteomyelitis (24), acute hip arthritis (63), and retinal vasculitis (51).

G. vaginalis-associated BV follows a variable clinical course but is prone to therapeutic failure, leading to persistence or recurrence and the formation of metronidazole-resistant biofilm infections (67). The ecological paradox of *G. vaginalis* being associated with both asymptomatic commensalism and BV could be explained by genotypic differences among strains that result in substantially different clinical phenotypes. Attempts to classify *G. vaginalis* strains based on laboratory phenotype have not been shown to be clinically relevant (3, 49, 74). Consequently, and because no closely related organisms are known to exist, surveys of *G. vaginalis* prevalence among the general population do not distinguish between those bacteria that have high 16S rRNA sequence identity to the *G. vaginalis* type strain (ATCC 14018; GenBank accession no. M58744.1) and those that do not.

These single-gene comparison studies and laboratory pheno-

Received 16 January 2012 Accepted 13 May 2012
Published ahead of print 18 May 2012
Address correspondence to Garth D. Ehrlich, gehrlich@wpahs.org, or Fen Z. Hu, fhu@wpahs.org.
* Present address: Adam Retchless, University of California, Berkeley, California, USA; Robert Boissy, University of Nebraska Medical Center, Omaha, Nebraska, USA.
A.A., J.E., and A.R. contributed equally to this article.
Supplemental material for this article may be found at http://jb.asm.org/.

doi:10.1128/JB.00056-12

typing assays are not capable of monitoring the large number of genes whose presence or absence can play a major role in determining bacterial phenotype (47). Genic content differences among *G. vaginalis* strains may underlie the diverse pathological features, outcomes, and sequelae that have been associated with this species. Thus, it is important to identify the gene possession differences that may be responsible for the production of particular clinical phenotypes in order to take appropriate measures to prevent adverse health outcomes.

Two prior genomic studies have approached this problem by comparing an isolate from a diseased patient to an isolate from an asymptomatic carrier of *G. vaginalis* (29, 78). Both studies identified differences in gene content, and yet the link between these differences and their pathogenic potential remains speculative, in part due to the fact that a pathogen may be present even in individuals who are not diseased, as is often seen in nasopharyngeal pathogens such as *Haemophilus influenzae* (15, 23, 43, 50, 77) and *Streptococcus pneumoniae* (60, 61). In addition, individuals are often found to carry multiple strains of *G. vaginalis*, any of which might be responsible for pathogenicity (11). To complicate matters further, carriers may remain asymptomatic even in cases of a dense colonization with *G. vaginalis*, as indicated by a high Nugent score (53). All of these factors, combined with a very incomplete understanding of the species-level supragenome and the gene possession differences among strains, make it difficult to understand the correlation between *G. vaginalis* genic content and pathogenicity. Therefore, to understand the genetic diversity underlying the virulence properties of *G. vaginalis*, a more complete characterization of its genetic potential is necessary.

The distributed genome hypothesis (DGH) states that bacterial pathogens associated with chronic infection are able to quickly adapt to changing conditions (e.g., nutritional shift, polyclonal infection, host immune response, antibiotic therapy, etc.) by acquiring novel genes from conspecifics (16, 17, 19). Phylogenetic modeling has shown that strains within the pateurellaceae lacking functional *com* genes are most likely of recent origin. The lack of ancient strains without a functional *com* regulon suggests that there exists selective pressure for the maintenance of transformation-promoting genes (58). Thus, genes involved in DNA uptake can be thought of as "population-level virulence" factors (35, 37). This hypothesis helps explain the fact that independent isolates of a bacterial species share a core set of genes but that many of the genes of a species are distributed only in subsets of individual strain genomes (16, 17, 19, 35, 70, 71). The DGH posits that pathogens that establish chronic polyclonal infections have strain-specific subsets of distributed genes that augment the species-defined core genome and that continual admixture of the distributed genes among strains creates a species-wide "supragenome" (or "pan-genome") which serves as an evolutionary strategy to maximize the population fitness of the species under diverse environmental conditions. Modern high-throughput genome-sequencing and analysis techniques now permit the direct investigation of horizontal gene transfer (HGT) during polyclonal infections (18, 31). One of the tenets of the DGH is that strains of a species exchange DNA via one or more HGT mechanisms; if HGT is not observed among strains, i.e., if they are evolving independently of one another due to a barrier for gene exchange, then it suggests either that ongoing speciation is present or that the strains involved belong to separate species.

MATERIALS AND METHODS

Strain acquisition, culture, and DNA preparation. All *G. vaginalis* strains sequenced for this study were clinical isolates obtained from the Magee-Womens Research Institute. These isolates produced betahemolytic reactions on human blood Tween (HBT) bilayer agar (Becton, Dickinson, Franklin Lakes, NJ) when incubated at 37°C in 6% CO_2 (72). All isolates were catalase-negative, gram-variable rods. Several have been described previously (49). Biotyping assays were conducted as previously described (49).

A single colony of each strain was used to prepare an inoculum in 1× phosphate-buffered saline (PBS) for lawn growth on HBT agar plates, as these strains form robust biofilms but tend to culture poorly in liquid media. Each strain was spread on 4 plates and incubated for 30 to 40 h as described above, after which the lawns for each strain were scraped together into PBS and the bacteria subjected to pellet formation by centrifugation at 5,500 × *g*. The bacteria were resuspended in TE buffer (10 mM Tris [pH 7.2], 1 mM Na_2EDTA) and lysed by the addition of 10% sodium dodecyl sulfate and RNase at 37°C for 1 h; proteinase K was then added and incubation continued at 55°C for 1 h followed by the sequential addition of NaCl, alkyltrimethyl-ammonium bromide (CTAB), and a 24:1 mixture of $CHCl_3$:isoamyl alcohol. After mixing and centrifugation (6,000 × *g*), the aqueous phase was removed and the DNA precipitated with isopropanol and then centrifuged and washed 3 times with 70% ethanol. The recovered high-molecular-weight DNA was then quantified using $UV_{A260/280}$ absorption spectroscopy and agarose gel electrophoresis; quality was evaluated using an Agilent Bioanalyzer.

To test for the presence of plasmids, strains were lawn grown and scraped into PBS as described above, plasmid DNA preparations were performed using a Qiaprep Spin minikit (Qiagen, Germantown, MD), and the samples were run on a 1% agarose gel.

Genomic sequencing. All sequences generated at the Center for Genomic Sciences (CGS) were obtained using Roche/454 Life Sciences GS FLX Titanium sequencing technology. Individual fragment libraries were prepared from the DNA of each of the 12 *G. vaginalis* strains to be sequenced as described in the *GS FLX Titanium, General Library Preparation Method Manual* (October 2008; Roche Molecular Systems, Nutley, NJ). Fragment library binding to beads, titration, emulsion PCR, emulsion breaking, bead enrichment, and picotiter plate-based pyrosequencing were performed as described in the *GS FLX Titanium emPCR and Sequencing Protocols* (October 2008). The genomes for two strains, B473 (7571-2) and B475 (65/20LIT), were completely closed using PCR-based primer walking, Sanger sequencing, and an ABI 3730xl sequencer.

Genome assembly and annotation. Following pyrosequencing, the number of reads, the average read lengths, and depth of coverage for each strain were determined. The raw sequence reads for each strain were assembled into contigs by the use of a Roche/454 Life Sciences GS *de novo* Newbler assembler (version 2.0.00.20 or 2.0.01.14) and the default parameters except for minimum overlap identity, which was adjusted to obtain the fewest contigs. The assembled genomes were submitted to RAST (Rapid Annotation using Subsystems Technology; http://RAST.nmpdr.org) for automated annotation (5). Specific gene functions were inferred to be present in a genome if any region of a genome was inferred to be homologous to a region in another genome (see below) with the annotated function (34).

Genome comparisons. The CGS comparative-genomics pipeline (8, 12, 14, 31, 32, 34) was used to identify homologous sequences among the genomes based on FASTA similarity statistics (55). Annotated coding sequences (CDS) were clustered on the basis that all genes in the cluster must be connected by a network of good matches (single linkage/70% amino acid [aa] identity/70% length); however, for some analyses, these criteria were relaxed (see below). Nonannotated homologs were identified by a search of contig sequences for regions that matched annotated genes with at least 70% nucleotide identity over 70% of the length of the annotated gene. The result is a set of gene clusters each of which is either present or absent in each genome and may have multiple representatives from a

TABLE 1 Clinical and phenotypic characteristics of the *G. vaginalis* strains used for whole-genome sequencing[a]

Clinical isolate	Clinical ID	Biotype	Nugent score	STI	Source	Symptoms/diagnosis
B472	0284V	1	7	*Chlamydia trachomatis*	Endometrium	Abnormal discharge, odor
B473	7571-2	1	1	Negative	Vagina	BV
B474	0288E	1	8	Negative	Endometrium	Abnormal discharge, odor
B475	64/20LIT	2	3	Negative	Endometrium	None
B476	64/20B	2	3	Negative	Endometrium	None
B477	55/15-2	3	8	Negative	Endometrium	None
B478	1400E	4	9	Negative	Endometrium	UNK
B479	1500E	5	7	Negative	Endometrium	UNK
B513	$007/03B_{MASH}$	2 or 5[b]	7	HSV-2	Vagina	BV
B482	$007/03C2_{MASH}$	2 or 5[b]	10	HSV-2	Vagina	BV
B483	$007/03D_{MASH}$	3 or 7[b]	3	HSV-2	Vagina	BV
B512	61/19V5	7	5	Negative	Vagina	None

[a] All strains were obtained from the Magee-Womens Research Institute in Pittsburgh, PA. Isolates B475 (64/20LIT) and B476 (64/20B) were simultaneously collected from the same patient. The three MASH strains were sequential isolates from the same patient, who was undergoing metronidazole treatment for bacterial vaginosis (BV). The Nugent score is an ecological assessment of bacterial flora based on a Gram-stained smear of vaginal discharge; scores of 7 and above are consistent with BV (53). ID, identifier; HSV-2, herpes simplex virus 2; STI, sexually transmitted infection; UNK, unknown.

[b] Lipase activity unknown.

given genome; clusters represented in all genomes are designated "core," and those remaining are designated "distributed."

NG. Two neighbor-grouping (NG) methods were used to examine the relationships of the strains (27). The first is based on the similarity of phylogenetically informative gene content (PIGC), a modification of the fraction of distributed genes (FDG) statistic (32), and the second is based on average nucleotide identity (ANI) (64). Each statistic has a value between zero and one, and the distance statistic was reported as 1 − ANI or 1 − PIGC. ANI was calculated using the genes found in all genomes and nucleotide FASTA alignment (55) of each annotated CDS against each contig from each strain. PIGC was calculated by identifying all distributed gene clusters whose presence can be considered phylogenetically informative (i.e., present in at least two genomes and absent in at least two genomes). By limiting the calculations to phylogenetically informative gene clusters, this statistic places a greater emphasis on the relatedness of genomes (whether arising from clonal ancestry or gene transfer) than the FDG does. Furthermore, it focuses on the gene presence polymorphisms that are most likely to be biologically important, since genes can be absent from a single genome due to incomplete sequencing (particularly for the previously published genomes that are in many contigs), and the annotated CDS that are unique to a single genome are unlikely to be maintained in the bacterial genome by selection (75).

Phylogenetics. All phylogenetic analyses were based on a set of 473 high-confidence multiple sequence alignments (MSAs), wherein each alignment contained one gene from each genome. These gene sets were constructed using the CGS comparative genomics pipeline (described above) and a 70% amino acid identity threshold for clustering MSAs. A gene was discarded if its constituent sequence was not aligned with each other sequence in the cluster over at least 70% of its length (i.e., minimum sequence overlap ≥ 70%). *G. vaginalis* is a member of the *Bifidobacteriaceae* (73); thus, to construct a rooted phylogeny based on single-copy core genes of the *G. vaginalis* strains, we first identified the gene clusters that contained one protein-encoding gene from each genome and then used BLAST to identify outgroup genes in each of four other *Bifidobacteriaceae* genomes (*Bifidobacterium bifidum* PRL2010 [NC_014638.1; submitted to NCBI 2 November 2010], *B. animalis* subsp. *lactis* DSM 10140 [NC_012815.1; 23 March 2011], *Scardovia inopinata* F0304 [NZ_ADCX00000000; 30 March 2010], and *Parascardovia denticolens* F0305 [NZ_ADEB00000000; 30 March 2010]) from translated lists of "cDNA" downloaded from the NCBI Genome website. If a single *G. vaginalis* core gene was identified in each outgroup genome, each gene being the best reciprocal BLAST hit to each *G. vaginalis* gene in the cluster, then the translated protein-encoding sequences were aligned using MAFFT (38) (version 6.489b; parameters "-maxiterate 1000 -geneafpair"). Alignments were excluded from further analysis if any two sequences within the alignment were not aligned over at least 70% of the length of each sequence. These alignments always had a mean residue pair (MRP) score > 0.87 according to the cooptimal multiple sequence alignment algorithm (45) (Perl script "COSv2.03.pl" provided by Giddy Landan). This resulted in 332 alignments which were back-translated to codon alignments, concatenated, and used for phylogeny reconstruction with PhyML 3.0.1 software (25) in which we employed an HKY substitution model in which the substitution rate corresponds to the equilibrium frequency of each target nucleotide and which uses empirically determined rates for transitions and transversions along with optimization of topology, branch length, and substitution parameters. (Phylogenies for individual core gene clusters were similarly constructed.) Congruence between gene trees and the whole-genome reference tree was evaluated for each reference clade individually by evaluating a rooted subtree of the gene tree, containing only the taxa present in the reference clade and its sibling clade in the reference tree. If any of these taxa were missing from the gene tree, the reference clade was not evaluated with that gene tree. The reference clade was accepted by the gene tree if it was monophyletic with strong bootstrap support (e.g., in excess of 190/200) in that subtree; the reference clade was rejected if there was strong bootstrap support for any clade that included a member of the sibling clade along with a member of the reference clade while also excluding a member of the reference clade. If monophyly was accepted or rejected with weak bootstrap support, the gene was not counted. If the parent clade was not monophyletic within the gene tree, then the subtree was rooted on the node analogous to the deepest possible node on the reference tree. Algorithms were implemented in Perl 5.8, using the BioPerl 1.6.1 libraries (66). NeighborNet diagrams were generated by SplitsTree4 (36).

Nucleotide sequence accession numbers. All genomic sequences have been deposited in GenBank (see accession numbers in Table 2).

RESULTS

Genome sequencing and assembly of diverse *G. vaginalis* clinical isolates. Twelve clinical isolates of *G. vaginalis* were selected to maximize heterogeneity of multiple factors, including clinical site of isolation (vagina and endometrium); comorbid conditions; patient symptoms; biotype; and Nugent score (Table 1). Whole-genome shotgun sequencing was performed on each of these strains to an average coverage depth of 41× (range, 25× to 63×), with average read lengths of 356 bases, using the 454 Life Sciences FLX Titanium platform (Table 2). Applying Lander-Waterman statistics (46), this sequencing depth is predicted to provide ≫99.99% coverage of each genome sequenced. Genomes were assembled by supplying the raw reads to the *de novo* assembler of

TABLE 2 Sequencing data for the 12 new and 5 previously sequenced *G. vaginalis* clinical strains[a]

Genome sequence ID	Clinical ID	Depth	Size (Mbp)	%GC	No. of Newbler contigs	No. of current contigs	Avg read length	Q39 minus bases (%)	NCBI GPID	NCBI accession no.
B472	284V	32	1.650	41.19	16	9	305	0.08	42431	NZ_ADEL00000000
B473	7571-2	32	1.672	41.28	9	1	295	0.07	42435	NZ_ADEM00000000
B474	0288E	38	1.709	41.23	17	17	327	0.07	42437	NZ_ADEN00000000
B475	64/20LIT	42	1.493	42.17	11	1	334	0.07	42439	NZ_ADEO00000000
B476	64/20B	45	1.493	42.19	14	14	365	0.06	42441	NZ_ADEP00000000
B477	55/15-2	34	1.643	41.29	25	25	351	0.09	42443	NZ_ADEQ00000000
B478	1400E	39	1.716	41.18	28	28	384	0.06	42445	NZ_ADER00000000
B479	1500E	49	1.548	42.96	27	27	382	0.05	42447	NZ_ADES00000000
B513	$007/03B_{MASH}$	49	1.567	42.27	16	16	394	0.03	42449	NZ_ADET00000000
B482	$007/03C2_{MASH}$	40	1.546	42.27	22	22	355	0.09	42451	NZ_ADEU00000000
B483	$007/03D_{MASH}$	63	1.491	43.40	11	7	384	0.03	42453	NZ_ADEV00000000
B512	61/19V5	25	1.501	43.26	12	12	394	0.04	42455	NZ_ADEW00000000
AMD[b]	NA	NA	1.606	42.08	NA	117	NA	NA	40893	NZ_ADAM00000000
5-1[b]	NA	NA	1.673	42.04	NA	94	NA	NA	40895	NZ_ADAN00000000
14018[78]	594	NA	1.603	41.19	NA	145	NA	NA	46675	NZ_ADNB00000000
14019[78]	317	NA	1.667	41.36	NA	1	NA	NA	31473	NC_014644
409-05[78]	NA	NA	1.617	42.02	NA	1	NA	NA	43211	NC_013721

[a] The first 12 sequences were newly determined during this study. Depth = depth of sequencing coverage for the genome; Size = the size of the genome in megabase pairs; %GC = the percentage of guanosine:cytosine base pairs for the genome; No. of Newbler contigs = the number of contigs generated from the raw sequence reads by the Roche/454 Life Sciences automated *de novo* assembler; No. of current contigs = the state of closure of the genome as of submission; Avg read length = the average number of bases read for each sequencing reaction by the sequencer for each genome; Q39 minus bases (%) = a sequence confidence metric which gives the percentage of bases where the error rate is expected to exceed 1 in 10,000. ID, identifier; GPID, NCBI's Genome Project identifier; NA, not available.

[b] Sequenced by the Center for the Study of Biological Complexity, Virginia Commonwealth University.

Newbler, which produced a mean of 17 contigs (range, 9 to 28) per genome. Visual inspection indicated that most of the assembly gaps were due to repeat sequences in the genome. Subsequently, using standard PCR-Sanger sequencing-based gap closure approaches, we closed the genomes of two of the isolates, B473 and B475. Each of the genomes of these two strains contains a single chromosome with no evidence of episomal elements, as was observed for the previously published strains 14019 and 409-05 (78).

We were not able to isolate plasmids from any of the other 10 sequenced strains in spite of repeated attempts (data not shown). All contigs in the 10 unclosed genomes had nucleotide composition (percent GC [%GC]) and sequencing depths that were characteristic of their genomes. An alignment of these contigs against the four closed genomes, made with Mauve Aligner using default parameters (59), identified several contigs that did not correspond to any segment of the closed genomes; however, most of these unaligned contigs did align with contigs from other unclosed genomes in the study. None of the nonaligning contigs contained sequences that were homologous to known plasmid sequences, as tested with a BLAST search of GenBank, suggesting that episomal elements are rare or absent among *G. vaginalis* strains. All genomic sequences have been deposited with GenBank (see accession numbers in Table 2).

Extensive genomic diversity among *G. vaginalis* isolates. A broad overview of the relatedness of these 17 genomes is presented by a NeighborNet diagram (10) (Fig. 1) constructed from the concatenation of 473 high-confidence core gene multiple-sequence alignments (MSAs; see Materials and Methods). The NeighborNet algorithm is similar to the neighbor-joining tree-building algorithm, but it also indicates where the distance matrix is not additive—possibly resulting from recombination between sequences—and presents this as a network with sequences represented by external nodes. Sequence divergence without recombination ideally results in a tree-like structure in the network, whereas recombination is expected to produce either star-like patterns (where all sequences are equally distant from each other) or reticulation (where one sequence is represented as a neighbor to two other sequences, despite those two sequences' genomes being distant from each other overall). The NeighborNet diagram suggests that there are two sets of strains (A and B) each composed of two groups (groups 1 and 2 and groups 3 and 4, respectively) within our *G. vaginalis* sample (color coded in Fig. 1). There is substantial divergence and very limited recombination between and among these groups, as indicated by the long branches separating the groups. However, two of the groups (groups 1 and 4—represented by green and red labels, respectively) have branching patterns consistent with recombination within the group. All strains in all four groups were categorized as *G. vaginalis* on the basis of their being isolated from the human female reproductive tract, laboratory phenotype, and high similarity (~98%) of 16S rRNA sequences to the *G. vaginalis* type strain (ATCC 14018) sequences. Interestingly, a phylogenetic tree of the 16S sequences also strongly supports the same four-group structure with slightly different topology (see Fig. S1 in the supplemental material).

Interspecies comparative genomic analyses revealed that the level of diversity among the 17 *G. vaginalis* strains was exceptionally high for a single bacterial species. This was determined by comparing the gene content diversity within *G. vaginalis* to that of 22 other bacterial species for which we could identify eight or more high-quality draft or complete whole-genome sequences (Table 3). The *G. vaginalis* diversity is directly apparent from the summary data of the individual genomes; for instance, chromosome sizes ranged from 1.491 to 1.716 Mb, and chromosomal GC content ranged from 41.18% to 43.40% (Table 2); the variance in the latter parameter is greater than what has been observed for any of the other 22 species and three times the average variance (data not shown). Two of the 17 *G. vaginalis* genomes were excluded

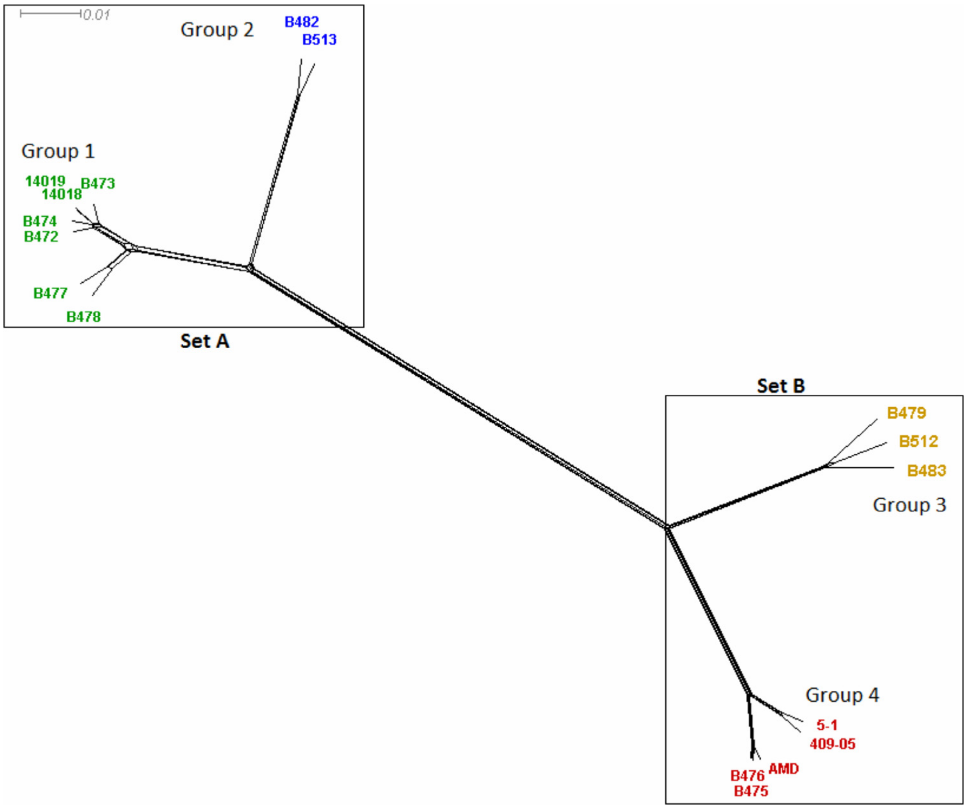

FIG 1 NeighborNet of 473 protein-encoding core genes, found in all 17 *G. vaginalis* genomes, aligned with high confidence. Two major sets of strains are apparent, A and B, each of which is composed of two groups, indicated by the coloring of the labels identifying each strain: group 1, green; group 2, blue; group 4, red; group 3, orange.

from some of the analyses, because they could not be considered to represent independent isolates. These duplicate genomes were unclosed, and each was essentially identical to a closed genome (Table 4, gray boxes). Genomes B475 and B476 were isolated from the same patient at the same time, and those genomes are essentially identical, so we excluded B476. Likewise, strains 14019 and 14018 were submitted to the American Type Culture Collection by the same researchers (20) without documentation that they originated from separate patients; the primary difference between these two genomes is that the sequence for 14018 lacked 46 genes that were found in 14019, but this is likely an artifact of the 14018 sequence being fragmented into 145 contigs, meaning that these genes could exist in the unsequenced regions.

Using our standard gene clustering parameters of 70% amino acid identity over at least 70% of the shorter sequence (34), which we have used for supragenomic analyses on more than two dozen other bacterial species, we identified only 746 core genes of 2,792 genes in the *G. vaginalis* supragenome (Table 3, line 1). Annotations for a member from each core gene cluster can be found in Table S1 in the supplemental material. Not only does *G. vaginalis* display the smallest core genome of all species examined in terms of the absolute number of gene clusters, but it also has the smallest percentage of each genome that is core and the smallest percentage of the species supragenome that is core. The large size of the *G. vaginalis* supragenome relative to its core genome does not appear to be the consequence of an extremely high rate of new genes being acquired from outside the species, since individual strains had on average only 21 unique genes, representing just 1.6% of each genome on average (Table 5) (gene annotations are listed in Table S3 in the supplemental material), which is one of the lowest values seen for any bacterial species.

Even when we lowered the threshold for clustering genes to 50% amino acid sequence identity for the *G. vaginalis* strains (to ensure that high allelic divergence was not interpreted as representing differences in gene possession), we could still identify only 894 core genes of a total of 2,195 (Table 3, line 6), which still results in the smallest core genome by count and the third-smallest core genome by percentage compared to all other species core genomes (that were each calculated using the 70% identity threshold). Even using these asymmetric criteria, only *Escherichia coli* and *Bacillus cereus* (both of which have genome sizes 4 to 5 times larger) have a lower percentage of core genes.

We also observed that, even using this lowered stringency criterion for clustering, there were many gene clusters that were present in only one or the other of the two major strain sets (i.e., present in either set A, composed of nine genomes, or set B, composed of eight genomes). This indicates that each of these major sets of strains contains an expanded core genome that is specific to that set (Fig. 2a). This type of core genome gene frequency behavior is absent from genome comparisons within all other bacterial species examined, including *Staphylococcus aureus* (Fig. 2b) and even those with very large distributed genomes such as *Bacillus cereus* and *Escherichia coli* (see Fig. S2A and B in the supplemental material), and is seen only in mixed-species data sets such as a

TABLE 3 Genomic diversity among bacterial species[a]

		Gene cluster count			Core genes as % of		% GC of core genes		
	Strain#	core	total	genome mean	total	genome mean	minimum	maximum	difference
Gardnerella vaginalis	**15**	**746**	**2792**	**1445**	**27**	**52**	**44.06%**	**45.84%**	**1.78%**
Escherichia coli	34	3143	9009	4834	35	65	52.59%	53.03%	0.44%
Bacillus cereus	13	3729	10679	5541	35	67	36.42%	36.66%	0.24%
B. cereus & B. anthracis	30	3639	10797	5574	35	65	NC	NC	NC
Clostridium perfringens	9	2048	4976	2990	41	68	29.57%	29.83%	0.26%
***Gardnerella vaginalis*[1]**	**15**	**894**	**2195**	**1299**	**41**	**69**	**44.06%**	**45.84%**	**1.78%**
Salmonella enterica	30	3488	8117	4762	43	73	53.62%	54.30%	0.68%
***G. vaginalis* Set B**	**7**	**996**	**1975**	**1344**	**50**	**74**	**43.83%**	**45.23%**	**1.39%**
Haemophilus influenzae	24	1538	3100	1961	50	78	38.81%	38.93%	0.12%
Clostridium botulinum	8	2881	5246	3613	55	80	29.22%	29.34%	0.12%
Campylobacter jejuni	10	1400	2511	1706	56	85	31.18%	31.26%	0.08%
Streptococcus pneumoniae	32	1632	2808	1959	58	73	41.28%	41.55%	0.27%
Streptococcus agalactiae	8	1553	2662	1983	58	78	36.20%	36.30%	0.10%
***G. vaginalis* Set A**	**8**	**1100**	**1861**	**1415**	**59**	**78**	**43.03%**	**44.12%**	**1.09%**
Streptococcus pyogenes	12	1527	2469	1844	62	83	39.40%	39.44%	0.04%
Listeria monocytogenes	20	2475	3982	2911	62	85	38.88%	39.12%	0.24%
Clostridium difficile	10	3100	4927	3682	63	84	29.79%	29.93%	0.14%
***G. vaginalis* Group 4**	**4**	**1085**	**1676**	**1324**	**65**	**82**	**43.36%**	**43.40%**	**0.04%**
Burkholderia pseudomallei	14	5809	8810	6717	66	86	68.43%	68.86%	0.43%
Pseudomonas aeruginosa	9	4717	7062	5593	67	84	67.36%	67.89%	0.52%
***G. vaginalis* Group 1**	**6**	**1158**	**1650**	**1385**	**70**	**84**	**42.67%**	**42.82%**	**0.16%**
Staphylococcus aureus	19	2363	3251	2763	70	86	33.81%	33.91%	0.10%
Francisella tularensis	16	1446	1972	1711	73	85	33.22%	33.52%	0.29%
Yersinia pestis	16	3412	4518	4151	76	82	48.87%	48.96%	0.10%
Moraxella catarrhalis	10	1755	2383	2383	77	90	43.46%	43.49%	0.03%
Borrelia burgdorferi	13	1008	1286	1152	78	88	28.81%	29.12%	0.32%
Burkholderia mallei	10	4073	5157	4971	79	82	68.43%	68.58%	0.16%
***G. vaginalis* Group 3**	**3**	**1064**	**1338**	**1191**	**80**	**89**	**44.37%**	**44.69%**	**0.31%**
***G. vaginalis* Group 2**	**2**	**1191**	**1303**	**1247**	**91**	**96**	**43.42%**	**43.42%**	**0.00%**
Mycobacterium tuberculosis	16	3505	3841	3788	91	93	65.52%	66.00%	0.48%
Bacillus anthracis	17	5043	5549	5467	91	92	36.07%	36.13%	0.06%

[a] Species are sorted by the size of the core genome as a percentage of the total number of gene clusters in that species. Various groupings of *G. vaginalis* strains are highlighted and bold. Sets A and B resulting from the first iteration of the application of the neighbor-grouping (NG) method are highlighted in blue, and groups 1 to 4 resulting from the second iteration of the NG algorithm are highlighted in green, with groups 1 and 2 belonging to set A and groups 3 and 4 to set B. The phylogenetic clade structure for *G. vaginalis* is identical to the NG group structure. The species with the greatest diversity (other than *G. vaginalis*) by each criterion are highlighted in yellow. [1] *G. vaginalis* genes were clustered at 50% amino acid identity, while other species and subsets of *G. vaginalis* were clustered at 70%.

combination of *Streptococcus pneumoniae* and *S. mitis* (see Fig. S2C in the supplemental material).

Identification of four genotypic clusters within *G. vaginalis*. Two different population-level processes could hypothetically produce the extremely high variability in gene content observed among the *G. vaginalis* strains: frequent horizontal transfer of genes from a single large *G. vaginalis* distributed genome (i.e., all strains are exchanging DNA with one another) or divergence of genetically independent populations with independent gene gains and losses. Gene transfer confounds phylogenetic analysis and may result in fairly homogenous genome properties regardless of the vertical relationships among the bacterial strains (62). To test whether there is a robust structure of relatedness among these *G. vaginalis* strains, we applied the neighbor-grouping (NG) algorithm (8, 27) to both nucleotide and genic distance matrices (Table 4). This algorithm evaluates whether the distance between strains in each pair is smaller than the average distance for all pairs and creates single-linkage clusters of those genomes that are most similar (i.e., "neighbor group complexes"). By simply testing if it is possible to group genomes, this method avoids imposing preconceptions on the data, as can occur when tree-building methods are applied.

Application of the NG algorithm to the *G. vaginalis* genomic data set (using either nucleotide differences or gene content differences) robustly divides the isolates into the same two sets of strains identified by the NeighborNet algorithm whether using the 50% identity (Table 4) or 70% identity (Table 6) gene-clustering algorithm. Not only are the members of each neighbor group linked by a network of "valid neighbors" (distances that are more than 2 standard errors below the mean of all pairwise distances), but all pairs of isolates that have one member in each of the two sets of strains (i.e., one in set A and one in set B) are clearly not neighbors, since all distances between the strains in such pairs are more than 2 standard errors above the mean distance (Table 7). Extensive sequence divergence within the species as a whole is reflected in the average nucleotide difference (1 − ANI) between pairs of genomes. For the 894 core genes (identified using the 50% clustering parameter), the average nucleotide difference within strain pairs that are not in the same group often exceeded the ~6% divergence (Table 4, top) that corresponds to the traditional species cutoff of 70% DNA-DNA reassociation (40).

Using the relaxed 50% clustering data, there are 1,301 distributed gene clusters of which 979 are phylogenetically informative (i.e., each was present in at least two genomes and absent from at least two genomes). When we looked at these informative distributed genes in strain pairs composed of strains from the two major

TABLE 4 Percent differences between genomes determined using the neighbor-grouping algorithm[a]

	B472	B473	B474	14018	14019	B477	B478	B482	B513	B475	B476	AMD	5-1	409-05	B483	B479	B512
B472		1.68	1.60	1.55	1.57	3.90	4.09	12.60	12.55	22.49	22.52	22.46	22.30	22.26	22.31	22.30	22.28
B473	19		1.38	1.38	1.40	4.00	4.26	12.58	12.55	22.41	22.45	22.43	22.22	22.26	22.28	22.17	22.20
B474	17	16		1.38	1.38	4.13	4.27	12.58	12.62	22.47	22.51	22.44	22.26	22.26	22.37	22.30	22.31
14018	18	16	21		0.04	4.03	4.17	12.47	12.44	22.13	22.11	22.05	21.88	21.99	21.99	21.85	21.85
14019	16	14	19	4		4.04	4.19	12.55	12.51	22.13	22.14	22.08	21.91	21.92	22.06	21.95	21.96
B477	26	26	30	26	26		2.24	13.10	13.04	22.48	22.52	22.51	22.30	22.34	22.36	22.36	22.30
B478	29	24	23	28	27	23		12.86	12.86	22.49	22.54	22.51	22.32	22.38	22.22	22.27	22.28
B482	30	32	35	34	32	34	40		2.39	22.95	22.98	22.90	22.70	22.73	22.32	22.28	22.28
B513	32	32	36	35	32	33	38	11		22.97	23.00	22.95	22.66	22.71	22.34	22.25	22.29
B475	57	54	59	56	56	54	59	52	52		0.00	1.00	3.97	3.80	15.41	15.39	15.36
B476	57	54	59	56	55	54	59	52	51	1		1.00	3.96	3.81	15.43	15.39	15.40
AMD	60	57	55	58	58	58	55	56	55	16	16		4.04	3.87	15.39	15.39	15.39
5-1	59	56	54	56	57	62	58	55	54	33	33	27		1.43	14.86	15.07	15.01
409-05	56	56	57	54	55	55	59	51	50	29	28	31	18		15.02	15.14	15.12
B483	50	53	55	51	51	49	54	48	48	32	32	34	40	35		4.89	4.42
B479	51	52	55	50	50	51	56	48	50	32	32	34	39	34	21		4.03
B512	48	51	53	48	48	50	53	46	48	33	33	34	39	34	16	20	

[a] Above diagonal, percent nucleotides within core genes that are different between the strains in each individual *G. vaginalis* pair; below diagonal, percent phylogenetically informative distributed genes that are either present or absent in both genomes of each strain pair. Nucleotide differences were calculated on 746 core genes; gene possession differences were calculated on 979 phylogenetically informative distributed genes. The clades (A and B) identified by the neighbor-grouping algorithm (NGA) are delimited by the thick lines in the middle, with interclade pairwise genome comparisons represented by a stippled background in the upper right and lower left quadrants. The groups, identified by running a second iteration of the NGA independently on each of the clades, are represented by colored boxes as follows: group 1, green; group 2, blue; group 3, orange; group 4, red. Nonindependent isolates are indicated with solid gray shading.

strains sets (A and B), it was observed that all such strain pairs differed in possession of >46% of these genes (Table 6, bottom). Pairwise comparisons of distributed gene possession differences between strains in the two different strain sets (determined using the standard 70% clustering criteria) showed that the number of shared genes was often less than the number of genes not shared, leading to negative comparison scores (Tables 6 and 8). Such negative values have not been observed even for a single strain pair for any of the other bacterial species so examined (8, 12, 14, 32, and 34 and data not shown). The average *G. vaginalis* values for these comparative parameters across all 136 possible strain pairs are presented in Table 9.

TABLE 5 Unique genes found in *Gardnerella vaginalis* clinical isolates[a]

Genome sequence ID	Clinical ID	Total no. of gene clusters	No. of unique genes	Unique genes (% of total)
B472	284V	1,329	3	0.2
B473	7571-2	1,343	6	0.4
B474	0288E	1,384	8	0.5
B475	64/20LIT	1,199	7	0.5
B477	55/15-2	1,328	5	0.3
B478	1400E	1,392	17	1.2
B479	1500E	1,203	35	2.9
B513	007/03B$_{MASH}$	1,294	22	1.7
B482	007/03C2$_{MASH}$	1,267	19	1.5
B483	007/03D$_{MASH}$	1,195	14	1.2
B512	61/19V5	1,198	13	1.1
AMD	NA	1,295	38	2.9
5-1	NA	1,372	61	4.4
14019	317	1,325	9	0.6
409-05	NA	1,334	65	4.9

[a] ID, identifier; NA, not available.

Iterative use of the NG algorithm split each of the two major strain sets into two groups, after which it was impossible to confidently partition the groups any further, resulting in four final groups (each composed of the same strains identified by the NeighborNet analysis): group 1, consisting of B472, B473, B474, 14019, B477, and B478; group 2, consisting of B482 and B513; group 3, consisting of B512, B479, and B483; and group 4, consisting of B475, AMD, 409-05, and 5-1. Looking only at strain pairs within each of the four groups (colored boxes in Table 4), the average nucleotide difference for all such strain pairs was within the range for traditional species; however, looking at the distributed genes for strain pairs composed of members from different groups even within the same set (either A or B) showed that they all differed in possession of >30% of these noncore genes.

This set and group structure is also apparent in a neighbor-joining tree constructed from the PIGC distances (Fig. 3). Likewise, the divergence of the two major strain sets is evident in comparisons of the two individual set core genomes to the 15-strain core genome, resulting in highly significant increases in size (47.5% and 33% for sets A and B, respectively) (Table 3, blue-highlighted rows). Splitting each of these sets into groups 1 and 2 and groups 3 and 4 (green-highlighted rows) resulted in further core genome size increases of >7% for all groups relative to their parent sets. Similarly, there is a significant reduction in the supragenome size of each set, and again for each group, compared to the overall 15-strain supragenome. The percent increases in the sizes of the core genomes going from 15 strains to sets is characteristic of moving from a family-level view to a species-level view, and the increase associated with the second-level split is typical of going from a genus compilation to a species compilation (14). Finally, the resulting ratios of core genome size to supragenome size for the individual groups are more typical of the single-species

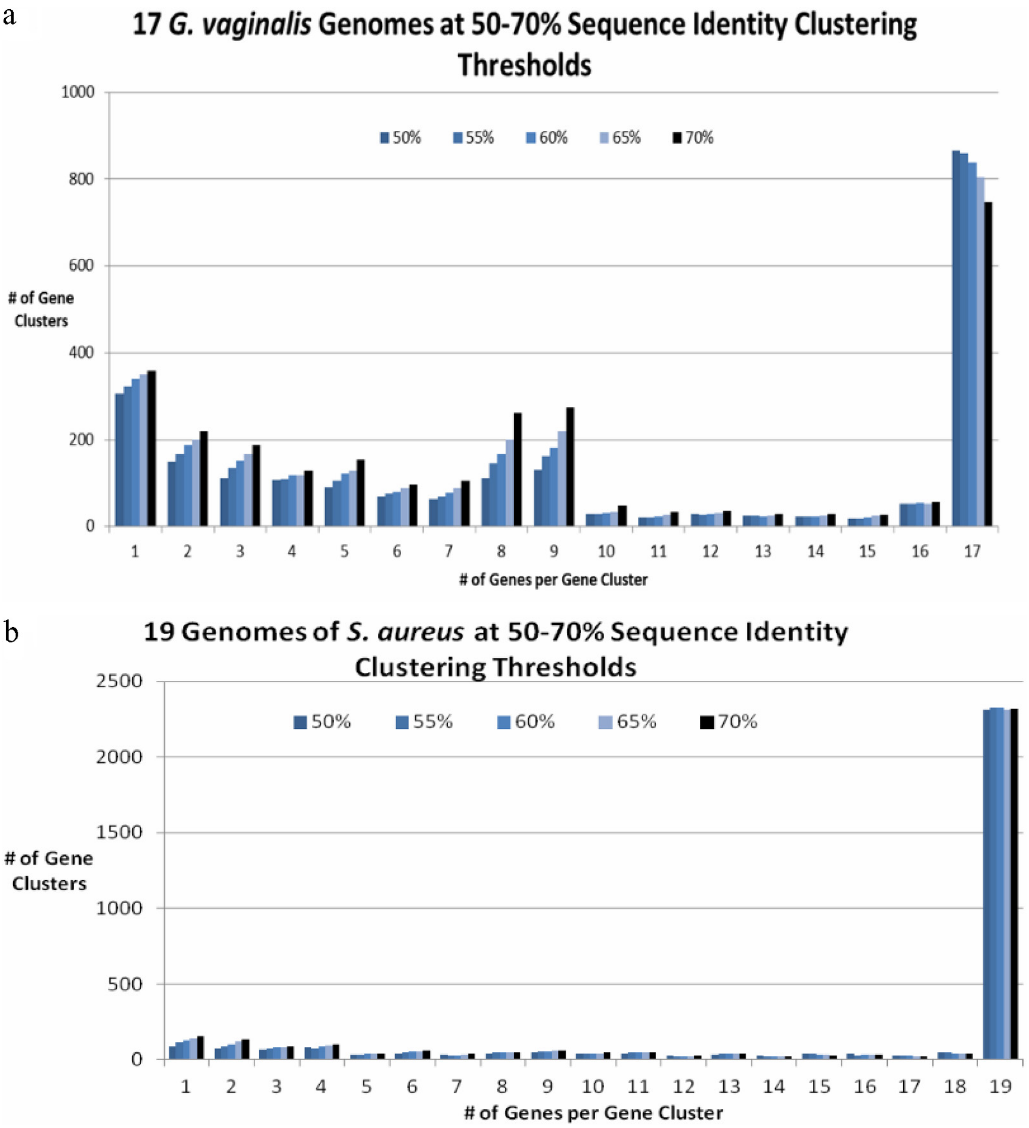

FIG 2 Histograms showing the dispersion of gene clusters among genomes, as a function of the clustering identity threshold, ranging from 50% to 70%. Horizontal axes depict the number of genomes for which the cluster has a representative, and the vertical axes count the number of clusters that are dispersed among the given number of genomes. The standard species threshold of 70% is represented by the black bars. (a) *Gardnerella vaginalis*. (b) *Staphylococcus aureus*. Additional species are described in Fig. S1 in the supplemental material.

data sets in our analysis than is the ratio for the entirety of the *G. vaginalis* population.

The independence of each of the four groups is also evident from examining their genomes at a global level. The range of genome sizes and GC contents within each group, compared to the 15-strain total, is much lower (Table 10), with the exception of the genome size for one strain (B475) in group 4 that is an outlier because of a major deletion (Table 2 and data not shown). It can be seen that the GC range for the entire strain assemblage is 2.22% (far greater than has ever been reported for a single species), but the greatest range for a group is 0.44% for group 3, which is typical of the diversity level seen in most species.

Phylogenetic examination for evidence of horizontal gene transfer among *G. vaginalis* strains. The neighbor-grouping results indicate that there is not a high level of horizontal transfer of distributed genes among the four *G. vaginalis* groups identified by both the NG and NeighborNet analyses. Likewise, the differences in nucleotide composition among the core genes of these four groups indicate that there is little recombination among the core genes. To test if these four groups—defined by differences in gene possession—are genetically isolated with respect to their core genomes, we examined the degree of congruence among the phylogenies of the core genes.

A reference phylogeny was constructed based on 332 core genes that exist as a single copy in each of the 17 *G. vaginalis* strains used in this study plus four *Bifidobacteriaceae* genomes used as out groups: *Bifidobacterium bifidum*; *B. animalis* subsp. *lactis*; *Scardovia inopinata* F0304; and *Parascardovia denticolens* F0305. These gene clusters were then back-translated to codon alignments, concatenated, and used for phylogenic reconstruction with PhyML 3.0.1 (25) (Fig. 4). The resulting topology had strong support at all nodes (SH [Shimodaira-Hasegawa]-like branch support = 1) and

TABLE 6 Gene comparisons between genomes[a]

	B472	B473	B474	14018	14019	B477	B478	B482	B513	B475	B476	AMD	5-1	409-05	B483	B479	B512
B472	0	1399	1436	1345	1408	1337	1351	1241	1253	878	877	895	926	926	909	929	931
B473	199	0	1434	1360	1419	1332	1381	1242	1254	910	909	923	950	935	899	922	922
B474	171	186	0	1352	1410	1333	1411	1237	1251	899	898	951	979	946	906	925	928
14018	223	204	266	0	1420	1279	1306	1178	1193	859	858	872	898	898	868	888	892
14019	163	152	216	66	0	1330	1359	1232	1248	896	895	908	932	932	905	925	929
B477	305	326	370	348	312	0	1385	1212	1228	898	897	913	920	934	924	924	920
B478	346	297	283	363	323	271	0	1212	1231	898	897	954	968	937	917	920	921
B482	429	438	494	482	440	480	549	0	1357	878	877	886	920	920	896	904	916
B513	440	449	501	487	443	483	546	157	0	896	895	912	938	940	914	917	928
B475	1092	1039	1107	1057	1049	1045	1114	1017	1016	0	1354	1306	1213	1222	1108	1119	1111
B476	1093	1040	1108	1058	1050	1046	1115	1018	1017	1	0	1306	1212	1221	1107	1119	1110
AMD	1160	1115	1105	1133	1127	1117	1104	1103	1086	200	199	0	1276	1235	1126	1144	1142
Gv5	1159	1122	1110	1142	1140	1164	1137	1096	1095	447	448	423	0	1346	1115	1123	1129
409	1129	1122	1146	1112	1110	1106	1169	1066	1061	399	400	475	314	0	1121	1137	1136
B483	999	1030	1062	1008	1000	962	1045	950	949	463	464	529	612	570	0	1202	1228
B479	1004	1029	1069	1013	1005	1007	1084	979	988	486	485	538	641	583	289	0	1229
B512	973	1002	1036	978	970	988	1055	928	939	475	476	515	602	558	210	253	0

[a] Above diagonal, number of gene clusters found in both genomes; below diagonal, number of gene clusters present in one genome but not the other. Following the clustering of genes into single linkage groups (minimum 70% amino acid identity over 70% of the smaller gene), the presence of each gene cluster was evaluated in each pair of genomes. Interclade comparisons are represented by a stippled gray background in the upper right and lower left quadrants. Neighbor-grouping categories are marked as follows: group 1, green; group 2, blue; group 3, orange; group 4, red. Nonindependent isolates are indicated with solid gray shading.

resulted in two major clades (clades A and B), each composed of two substituent clades (clades 1 and 2 and clades 3 and 4, respectively) whose topology exactly recapitulated the branching pattern of the groups identified by the NG method. Phylogenies constructed from higher-quality alignments (182 alignments with a minimum sequence overlap of 90%; MRP > 0.97) had the same topology but slightly shorter branch lengths, with support for the clade containing B473, 14019, and 14018 dropping to 95%.

To examine the extent of horizontal gene transfer between the major clades (A and B) and the four substituent clades (clades 1 to 4) revealed by the phylogenetic analyses, we constructed phylogenies for each individual gene cluster and counted the number of genes that strongly supported or rejected the monophyly of each clade relative to its sibling clade (bootstrap > 190/200 (Fig. 4, node labels). The clade A structure, containing clades 1 and 2, was rejected by only 7 of 227 alignments (i.e., those representing species that were likely to have experienced HGT); clade 1 was rejected by 21 of 234 alignments and clade 2 by only 5 of 300. These very low levels of rejection of the consensus topology could reflect either a small amount of HGT between clades or stochastic errors in the phylogenetic inference. Similar results were observed for clade B as a whole in that only 8/136 genes rejected the consensus topology. Looking at the subsequent branches, in clades 4 and 3, respectively, again only 12 of 304 and 14 of 240 gene phylogenies rejected the consensus topology. Thus, it appears that there is very limited HGT between clades. However, there is strong evidence of high levels of HGT within 3 of the 4 numbered clades: in clade 3, almost half of the alignments that produced a strong signal (95/205) rejected the consensus clade of B512 and B479; in clade 4, 41/303 gene trees rejected the 5-1/409-05 consensus clade; in clade 1, 85/273 genes rejected the B477/B478 branch; and also in clade 1, the B472 branch was rejected by 94/259 gene trees. Only in clade 2, consisting of only two strains, was there an appreciable lack of evidence of HGT.

TABLE 7 Statistical values obtained from the neighbor-grouping analyses[a]

	50% identity		70% identity	
Parameter	Genic	Allelic	Genic	Allelic
Mean distance	33.40	15.54	36.17	13.71
SE	0.79	0.50	1.06	0.44
2× SE above mean	34.98	16.54	38.29	14.59
2× SE below mean	31.81	14.54	34.05	12.83

[a] Values greater than twice the standard error above the mean indicate strain pairs that are not neighbors; values less than twice the standard error below the mean indicate that strain pairs are neighbors.

For Fig. 4, the analysis was limited to gene clusters with at least 80% minimum sequence overlap (MRP > 0.88), and only clades supported by at least 190/200 bootstrap replicates were evaluated. The ratio of support to rejection for each clade was robust with respect to modifications of the method, including raising the minimum sequence overlap to 90% (MRP > 0.95) and bootstrap support to 200/200. To include genes that did not have homologs in all genomes without introducing biases from the uneven distribution of genes, we tested each group using only those gene clusters that contained one sequence from each genome in both that group and the sibling group (i.e., representing a core cluster for the parent clade), and an outgroup sequence was present to allow rooting of the parent clade. If the parent clade was not monophyletic, then it was rooted on the node analogous to the deepest node on the strain phylogeny. This allowed us to test for recombination between the sibling groups without the confounding effect of different genes having different sets of possible topologies. Still, the results were robust with respect to variations of this requirement, including requiring that the parent clade be monophyletic; requiring the gene clusters to be present in all 21 genomes (i.e., *Bifido-*

TABLE 8 Pairwise gene possession comparison values for all possible *G. vaginalis* strain pairs[a]

	B472	B473	B474	14018	14019	B477	B478	B482	B513	B475	B476	AMD	Gv5	409	B483	B479	B512
B472		1200	1265	1122	1245	1032	1005	812	813	-214	-216	-265	-233	-203	-90	-75	-42
B473	1200		1248	1156	1267	1006	1084	804	805	-129	-131	-192	-172	-187	-131	-107	-80
B474	1265	1248		1086	1194	963	1128	743	750	-208	-210	-154	-131	-200	-156	-144	-108
14018	1122	1156	1086		1354	931	943	696	706	-198	-200	-261	-244	-214	-140	-125	-86
14019	1245	1267	1194	1354		1018	1036	792	805	-153	-155	-219	-208	-178	-95	-80	-41
B477	1032	1006	963	931	1018		1114	732	745	-147	-149	-204	-244	-172	-38	-83	-68
B478	1005	1084	1128	943	1036	1114		663	685	-216	-218	-150	-169	-232	-128	-164	-134
B482	812	804	743	696	792	732	663		1200	-139	-141	-217	-176	-146	-54	-75	-12
B513	813	805	750	706	805	745	685	1200		-120	-122	-174	-157	-121	-35	-71	-11
B475	-214	-129	-208	-198	-153	-147	-216	-139	-120		1353	1106	766	823	645	633	636
B476	-216	-131	-210	-200	-155	-149	-218	-141	-122	1353		1107	764	821	643	634	634
AMD	-265	-192	-154	-261	-219	-204	-150	-217	-174	1106	1107		853	760	597	606	627
Gv5	-233	-172	-131	-244	-208	-244	-169	-176	-157	766	764	853		1032	503	482	527
409	-203	-187	-200	-214	-178	-172	-232	-146	-121	823	821	760	1032		551	554	578
B483	-90	-131	-156	-140	-95	-38	-128	-54	-35	645	643	597	503	551		913	1018
B479	-75	-107	-144	-125	-80	-83	-164	-75	-71	633	634	606	482	554	913		976
B512	-42	-80	-108	-86	-41	-68	-134	-12	-11	636	634	627	527	578	1018	976	

[a] Comparison values were computed by subtracting the number of gene possession differences from the number of gene possession similarities for a given pair of strains. Interclade comparisons are represented by a stippled gray background in the upper right and lower left quadrants. Note that all comparisions between pairs in which one member is from clade A and one member is from clade B (upper right and lower left quadrants) have negative comparison scores, indicating that they share fewer than 50% of their genes. Neighbor-grouping categories are marked as follows: group 1, green; group 2, blue; group 4, red; group 3, orange. Nonindependent isolates are indicated with solid light gray shading.

bacteriaceae core genes); or allowing sequences from outside the parent clade to reject monophyly of the group being tested.

Some biotype markers correspond to genotypic clusters. All genomes in group/clade 1 were annotated as containing an operon with a *lacZ* gene (EC 3.2.1.23); a three-component ABC-type sugar transporter; and an α-L-fucosidase. Whereas these genes were core to clade 1, they were absent from all strains in the other three clades (see Table S2 in the supplemental material). The α-L-fucosidase gene in other bifidobacteriaceae has been associated with the ability to degrade both glycans and mucins (4), and it is therefore possible that this enzymatic activity provides for invasion through the mucosal layer, accounting for group/clade 1's association with endometritis (Table 1). Importantly, Turroni et al. have associated this enzymatic capability with foraging of host-derived glycans (73). Also core to this taxonomic grouping, but absent from the other three groups, are two operons containing transcriptional regulators associated with sugar metabolism: one is an NagC/XylR type and the other is a LacI type. These operons contain multiple other genes annotated as being involved in galactose and arabinose metabolism, suggesting that this clade has specialized by the acquisition of additional carbohydrate utilization mechanisms. The finding of the *lacZ* gene in this group is of particular interest in interpreting *G. vaginalis* diversity, since a β-galactosidase assay has traditionally been used to classify *G. vaginalis* isolates into biotypes (being positive in biotypes 1, 4, 6, and 8) (56). Our biotype data (Table 1) demonstrate that all isolates from these biotypes—except B477—tested positive on the β-galactosidase enzymatic assay, while no other isolate tested positive. For the isolates sequenced by others, ATCC 14018 is reported to be β-galactosidase positive (http://www.ncbi.nlm.nih.gov/protein/308234929), while the biotype of ATCC 14019 is not available. These facts indicate that a positive test for β-galactosidase activity may be an informative, but imperfect, test for the presence of these *lacZ* homologs and the associated genomic properties of the group/clade 1 *G. vaginalis*. Furthermore, fermentation of galactose and arabinose has been proposed by Benito et al. as an additional criterion for biotyping (6). In the population described in that study, fermentation of arabinose and galactose generally cooccurred with a positive test for β-galactosidase activity, a finding which our genomic data support.

Group/clade 2 is distinguished by the unique possession of a serine endopeptidase, ScpC, which is a recognized virulence factor in the streptococci that is associated with destruction of host chemokines, resulting in loss of signaling for polymorphonuclear cells (30). In addition, this group contains other proteases, including an HtpX homolog not found among the other three groups. Groups/clades 3 and 4 (clade B) both appear to contain EbpS, a cell surface elastin binding protein that is not found in the other major split. However, only in the clade 3 strains was it annotated as such. By using a BLASTX approach, we identified in clade 4 a core gene with similar structure by reducing the clustering stringency to 50% aa identity over 50% of the shorter sequence, as these genes did not cluster using our standard within-species clus-

TABLE 9 Average pairwise comparisons (number of gene clusters) among all possible *G. vaginalis* strain pairs

Parameter	Similarity	Difference	Comparison	Pair unique
Minimum	858.00	1.00	−265.00	0
Maximum	1,436.00	1,169.00	1,354.00	61.00
Avg	1,075.44	739.94	335.50	1.62
SD	187.38	358.50	541.48	6.74

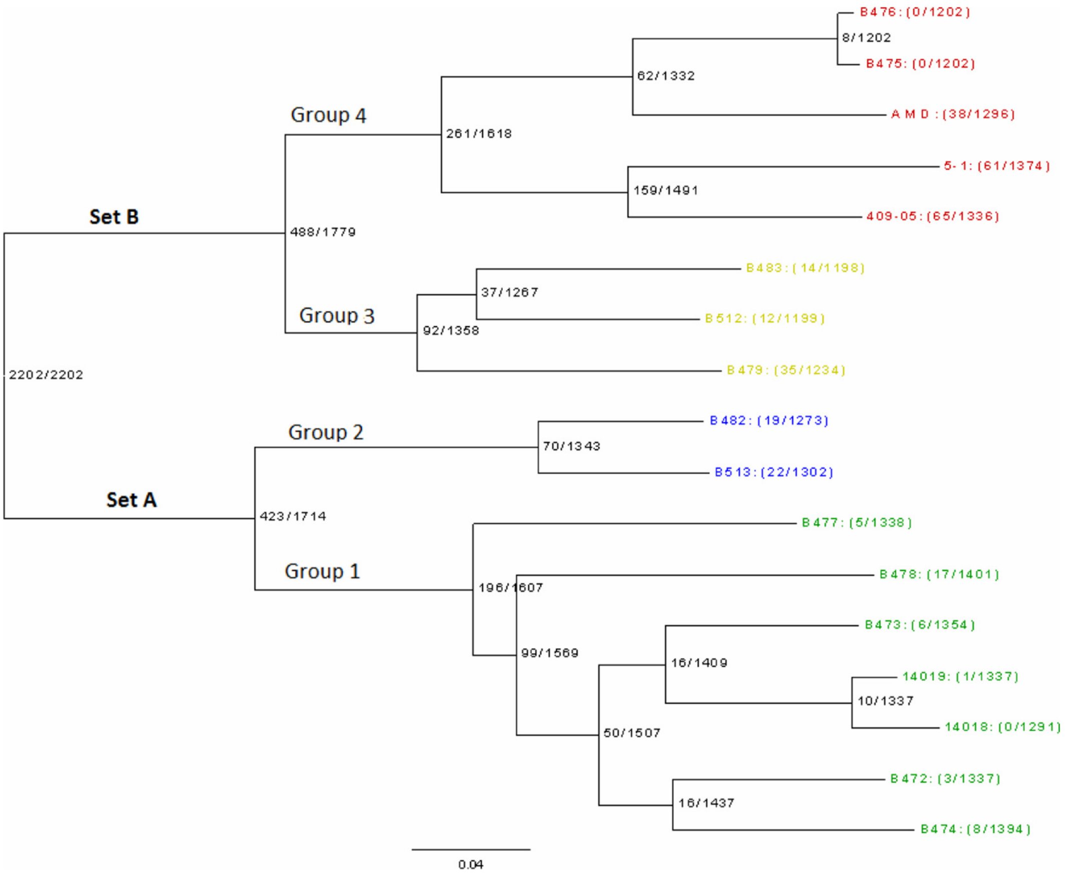

FIG 3 Neighbor-joining tree of 17 *G. vaginalis* genomes based on phylogenetically informative gene content; the midpoint is rooted at the left. The four neighbor-grouping complexes are colored differently. Node labels indicate the gene clusters found among the set of genomes associated with each node; the numerator is the number of gene clusters that are unique to that set, and the denominator is the total number of gene clusters with members in that set.

tering algorithm (34). This gene encodes an enzyme which in *Staphylococcus aureus* has been characterized as a virulence factor associated with intercellular matrix degradation of the host (54). Clade 4 also contains a unique core operon associated with allantoin utilization which may provide it with a unique scavenging/foraging ability absent in the other strains. Thus, each of the clades contain its own unique core genes which are predicted to provide either unique metabolic capabilities which could be associated with niche specialization or virulence factors providing for invasiveness. It is important that the vast majority of the clade-specific core genes for all clades are currently unannotated.

Vaginolysin shows evidence of recombination between groups. Recently, it has been reported that *G. vaginalis* possesses a member of the cholesterol-dependent cytolysin gene family, vaginolysin, that is species specific for human cells and encodes a pore-forming toxin that binds to the CD59 human complement regulatory molecule (63). It is believed that the action of this gene product along with proteolysis causes the breakdown of tissue and production of putrescine, resulting in the characteristic odor associated with BV. The *G. vaginalis* vaginolysin (*vly*) gene has been implicated as an acute virulence factor (21, 57), and we identified it as a core gene. However, this gene was distinctive as one of the few genes that rejected the broad phylogenetic relationships that we described above. To examine this more closely, we constructed a NeighborNet diagram based on the aligned amino acid sequences of this gene

TABLE 10 Group-specific ranges for genome size and GC content

Group[a]	Genome size range (Mb)	Maximum size difference (Mb)	Avg genome size (Mb)	%GC range	Maximum %GC difference	Avg %GC
All 15 strains	1.49–1.71	0.22	1.61	41.18–43.40	2.22	42.00
1	1.64–1.71	0.07	1.68	41.18–41.29	0.11	41.23
2	1.55–1.57	0.02	1.56	42.27	0.0	42.27
4	1.49–1.67	0.18	1.60	42.02–42.17	0.15	42.07
3	1.49–1.55	0.06	1.51	42.96–43.40	0.44	43.21

[a] Group 4 contains one outlier (B475) in terms of genome size; genome sizes of all other strains in this group are tightly clustered between 1.61 and 1.67 Mb.

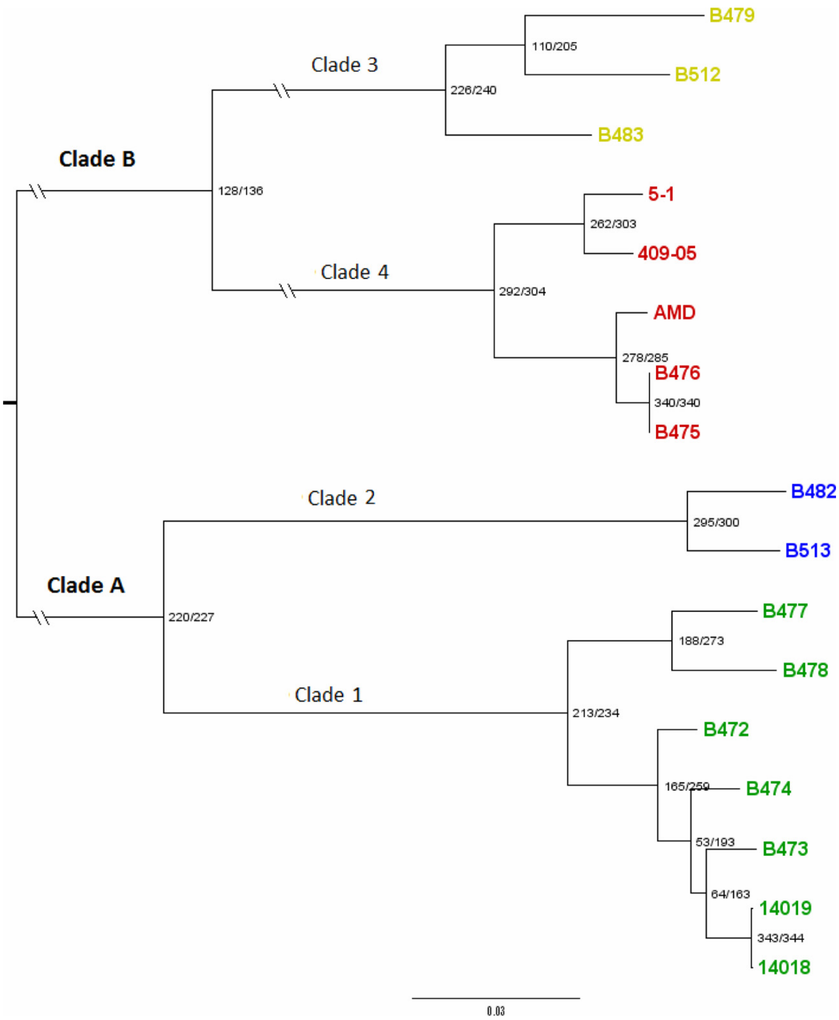

FIG 4 Maximum-likelihood phylogeny calculated for the concatenated alignment of 332 genes shared among all 17 *G. vaginalis* strains. All nodes have SH-like confidence values of 1. Node labels indicate the total number of gene clusters that confidently supported the clade (bootstrap > 190/200) relative to the total number that could confidently support or reject the clade.

(Fig. 5). The reticulate structure connecting sequences from groups/clades 1, 2, and 4 indicates a history of recombination among these genomes at this locus, with intragenic recombination breakpoints; only group/clade 4 retains its distinct identity at this locus. In addition to the split-tree analysis, a breakpoint analysis on the vaginolysin cluster was also performed using HyPhy version 2.0 (3, 41) software. Both AIC and AIC-c criteria showed a breakpoint at position 920 in the trimmed *vly* gene cluster alignment. Using the same software, the Kishino-Hasegawa test indicates that this breakpoint is likely due to recombination at a *P* value of 0.01.

Association of *G. vaginalis* groups with specific clinical conditions. Four of the five independent endometrial isolates belong to group/clade 1; however, the numbers are too small to provide statistical significance. This is also the group that contains the ß-galactosidase gene and the other genes associated with carbohydrate metabolism, raising the possibility that individual groups have proclivities for particular ecological niches within the human female reproductive tract.

DISCUSSION

It is clear that *G. vaginalis* as a species displays a wide range of clinical and metabolic phenotypes. Some *G. vaginalis* strains are members of the normal vaginal flora, where they are not associated with clue cells and BV, whereas others have been identified as major etiological agents of BV, where they are associated with overgrowth relative to the rest of the vaginal microflora, particularly the *Lactobacillus* sp. (7, 48). In addition, certain biotypes have been more highly associated with BV than others (9); however, this observation has been called into question (49). More recently, *G. vaginalis* has also been linked to endometritis and a number of otherwise sterile site infections, including cases of vertebral osteomyelitis with discitis, pelvic arthritis, and bacteremia (1, 44, 52, 65). Phenotypic variability has also been documented in terms of enzymatic function, with some strains characterized as producers or nonproducers of ß-galactosidase, sialidase, or lipase (49). Thus, it would appear that there exists within the *G. vaginalis* species a wide range of phenotypes, in terms of both metabolic capabilities and disease association. Due to its importance in normal vaginal

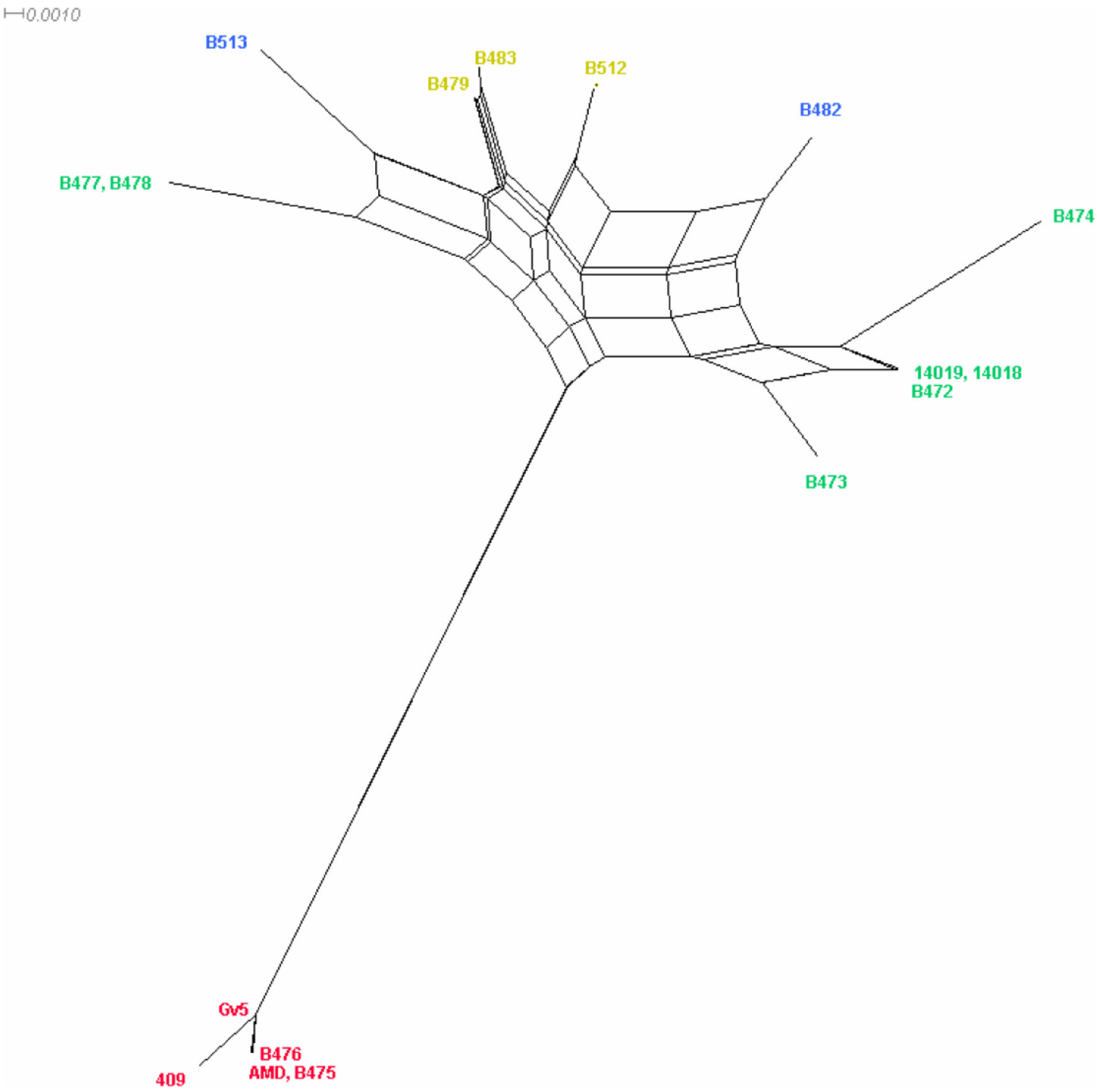

FIG 5 NeighborNet diagram derived from the vaginolysin (*vly*) gene sequences.

ecology, reproductive pathology, and invasive disease, we performed genomic and comparative genomic analyses on a diverse set of 17 strains to determine if there were significant allelic and gene content differences among the strains that could possibly account for the various observed phenotypes and to determine if there was a phylogenetic substructure present within the species.

If we start with the assumption that all the strains that are typed as *G. vaginalis* are a single species and then perform genome-wide comparisons among the strains, we are confronted with a number of unusual aspects compared with all other bacterial species that have been examined to date. First, it is observed that the %GC range of the core genes (1.78%) is more than 3 times greater than that of the next most diverse species, *Pseudomonas aeruginosa* (0.52%), which has an average genome size approximately four times that of *G. vaginalis* and is known to take up DNA from many other species (39). Next, it is observed that *G. vaginalis* possesses the smallest core genome, calculated by three different criteria, of any of the two dozen bacterial species for which there are eight or more high-quality draft or complete genomes available for comparative analyses. *G. vaginalis* has the lowest number of core genes (n = 746), with the next smallest core genome belonging to *B. burgdorferi*, consisting of 1,008 genes. The average percentage of each *G. vaginalis* genome that is core is also the lowest (51.6%), with *E. coli* having the next smallest percentage (65%). Finally, the *G. vaginalis* core genome is the smallest as a percentage of the supragenome (27%), with *E. coli* and *B. cereus* being the next most diverse, with 35% of their supragenomes being core. These observations are all the more surprising because the individual *G. vaginalis* genomes are very small (mean = 1.59 Mb), whereas all of the other species with the most highly variable genomes have genomes 3 to 4 times as large. Looking at another metric, the average gene possession differences for all possible pairs of *G. vaginalis* strains is 740; this is 81% to 150% higher than for all other bacterial species examined at this level of detail, including the highly recombinogenic *H. influenzae* (n = 395) (34), *Streptococcus pneumoniae* (n = 407) (32), and *Staphylococcus aureus* (n = 296) (8). Moreover, as a percentage of genome size, the *G. vaginalis* gene possession difference data are even more striking, as the average genomes of those other species are 17%, 32%, and 78% larger, respectively. In addition, *G. vaginalis* displays the highest degree of genomic plasticity in the pairwise comparison score, which is defined as the similarity score (genes in common) minus the difference score (genes not in common). Almost 50% of *G. vaginalis* strain pairs have a negative comparison score, meaning that they have less than 50% of their genes in common, and yet no other species has a negative comparison score for even a single strain pair. Surprisingly, in spite of the extremely high average degree of gene possession differences among the *G. vaginalis* strains, as a species it contains a relatively low percentage (9.8%) of unique genes (those present in only a single strain) compared to other species. For example, *H. influenzae*, which has a similar size genome, has unique gene complements that account for 17% of the species supragenome. This suggests that a high proportion of the *G. vaginalis* distributed genes are fixed in the population and may be providing important survival or virulence traits as opposed to be being orphan genes that are not providing the bacteria with an evolutionary advantage (70); however, we cannot rule out the possibilitiy that some of the genes that are important in host-pathogen interactions may be recently introduced orphan genes from other species that have not yet had time to become fixed in the population.

These diversity data, combined with the common clade and group structure supported by multiple independent analysis methods, including standard phylogenetic analyses, neighbor grouping, NeighborNet, genome size, and GC content, robustly demonstrate that there is very minimal HGT between groups/clades and provide a strong argument for separating the four clades/groups into individual species.

Core genome analyses can also be used to infer which strains form natural taxonomic groupings. Within a species, for example, the number of core genes decreases very slowly after the first several strains have been added during a supragenome analysis (8, 12, 14, 32, 34, 70). Thus, if the addition of a new strain results in a significant (>5%) decrease in the size of the core genome, it is suggestive that the newly added strain may belong to a separate species, as has been shown among the streptococci (14). Conversely, if the addition of a strain from what is thought of as a separate species to another species core genome does not result in a significant reduction in the core genome size of another related species, it is likely that the two species are actually one and the same, as has been observed for *B. cereus* and *B. anthracis* (Table 3, line 4). Supragenome analysis of *G. vaginalis* as a single species results in very large drops in the size of the core genome as different groups/clades are added; thus, by this high-level comparative genomic measure, it was also possible to divine the same substructure as was identified by the phylogenetic and multiple other similarity analyses discussed above.

The extraordinarily high degree of genomic plasticity observed among the 17 *G. vaginalis* strains evaluated in this study is supportive of the distributed genome hypothesis (8, 12, 16, 18, 19, 31, 32, 35) and suggestive of very extensive horizontal gene transfer (HGT) mechanisms being active, but only within the individual groups, as alternative mechanisms for the creation of such highly mosaic genomes are unknown. For HGT to produce such extensive mosaicism among the strains of a species, their colonizations and/or infections must be polyclonal in nature (at least in some cases). We do not have definitive proof of polyclonality at a single time point, but the fact that we were able to obtain three genomically distinct strains (B513, B482, and B483) obtained serially (14 May 2008, 16 June 2008, and 6 August 2008) from a single patient suffering from metronidazole-resistant BV certainly suggests that two or more strains may have been present simultaneously. Further, the observation that two of these strains (B513 and B482) are much more closely related than the average strain pair in terms of both distributed gene possession differences ($n = 125$ versus $\bar{x} = 608$) and the PairUnique parameter ($n = 68$; this value is 4 SD above the $\bar{x} = 4.77$) is consistent with, although not proof of, the hypothesis that one of them may be a progenitor of the other which evolved via one or more HGT events from one or more donors such as has been observed for pneumococcus (32). Future detailed gene-by-gene comparisons of these strains should be able to resolve this question.

These studies represent the first detailed genomic and comparative genomic analyses, both intraspecific and interspecific, performed for this important pathogen. Our observations support the hypothesis that individual strain differences, manifest in the realization that the current *G. vaginalis* species might more accurately be split into four species, may underlie the collective strains' association with a wide range of clinical conditions.

ACKNOWLEDGMENTS

This work was supported by Allegheny General Hospital, Allegheny Singer Research Institute, Magee-Womens Research Institute, a system usage grant from the Pittsburgh Supercomputing Center (G.D.E.), and NIH grants K23064396 (T.L.C.), U19AI084024 (S.L.H. and G.D.E.), DC04173 (G.D.E.), DC02148 (G.D.E.), DC02148-16S1 (G.D.E.), and AI080935 (G.D.E.).

We thank Mary O'Toole for help in the preparation of the manuscript.

REFERENCES

1. **Amsel R, et al.** 1983. Nonspecific vaginitis: diagnostic criteria and microbial and epidemiologic associations. Am. J. Med. **74**:14–22.
2. **Anukam KC, Osazuwa EO, Ahonkhai I, Reid GG.** 2006. Lactobacillus vaginal microbiota of women attending a reproductive health care service in Benin City, Nigeria. Sex. Transm. Dis. **33**:59–62.
3. **Aroutcheva AA, Simoes JA, Behbakht K, Faro S.** 2001. *Gardnerella vaginalis* isolate from patients with bacterial vaginosis and from patients with healthy vaginal ecosystems. Clin. Infect. Dis. **33**:1022–1027.
4. **Ashida H, et al.** 2009. Two distinct alpha-L-fucosidases from *Bifidobacterium bifidum* are essential for the utilization of fucosylated milk oligosaccharides and glycoconjugates. Glycobiology **19**:1010–1017.
5. **Aziz RK, et al.** 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9**:75. doi:10.1186/1471-2164-9-75.
6. **Benito R, Vazquez JA, Berron S, Fenoll A, Saez-Neito JA.** 1986. A modified scheme for biotyping *Gardnerella vaginalis*. J. Med. Microbiol. **21**:357–359.
7. **Biagi E, et al.** 2009. Quantitative variations in the vaginal bacterial population associated with asymptomatic infections: a real-time polymerase chain reaction study. Eur. J. Clin. Microbiol. Infect. Dis. **28**:281–285.
8. **Boissy R, et al.** 2011. Comparative supragenomic analyses among the pathogens *Staphylococcus aureus*, *Streptococcus pneumoniae*, and *Haemophilus influenzae* using a modification of the finite supragenome model. BMC Genomics **12**:187. doi:10.1186/1471-2164-12-187.
9. **Briselden AM, Hillier SL.** 1990. Longitudinal study of the biotypes of *Gardnerella vaginalis*. J. Clin. Microbiol. **28**:2761–2764.
10. **Bryant D, Moulton V.** 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. **21**:255–265.
11. **Cherpes TL, Hillier SL, Meyn LA, Busch JL, Krohn MA.** 2008. A delicate balance: risk factors for acquisition of bacterial vaginosis include sexual activity, absence of hydrogen peroxide-producing lactobacilli, black race, and positive herpes simplex virus type 2 serology. Sex. Transm. Dis. **35**:78–83.
12. **Davie JJ, et al.** 2011. Comparative analysis and supragenome modeling of

twelve *Moraxella catarrhalis* clinical isolates. BMC Genomics **12**:70. doi:10.1186/1471-2164-12-70.

13. **Demba E, Morison L, van der Loeff MS, Awasana AA, Gooding E, Bailey R, Mayaud P, West B.** 2005. Bacterial vaginosis, vaginal flora patterns and vaginal hygiene practices in patients presenting with vaginal discharge syndrome in The Gambia, West Africa. BMC Infect. Dis. **5**:12. doi:10.1186/1471-2334-5-12.
14. **Donati C, et al.** 2010. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol. **11**: R107. doi:10.1186/gb-2010-11-10-r107.
15. **Ecevit IZ, McCrea KW, Marrs CF, Gilsdorf JR.** 2005. Identification of new hmwA alleles from nontypeable *Haemophilus influenzae*. Infect. Immun. **73**:1221–1225.
16. **Ehrlich GD.** 2001. The biofilm and distributed genome paradigms provide a new theoretical structure for understanding chronic bacterial infections. Abstr. 41st Intersci. Conf. Antimicrob. Agents Chemother., Chicago, IL, p 524. American Society for Microbiology, Washington, DC.
17. **Ehrlich GD, et al.** 2010. The distributed genome hypothesis as a rubric for understanding evolution *in situ* during chronic infectious processes. FEMS Immunol. Med. Microbiol. **59**:269–279.
18. **Ehrlich GD, Hiller NL, Hu FZ.** 2008. What makes pathogens pathogenic. Genome Biol. **9**:225. doi:10.1186/gb-2008-9-6-225.
19. **Ehrlich GD, Hu FZ, Shen K, Stoodley P, Post JC.** 2005. Bacterial plurality as a general mechanism driving persistence in chronic infections. Clin. Orthop. Relat. Res. **437**:20–24.
20. **Gardner HL, Dukes CD.** 1959. *Hemophilus vaginalis* vaginitis. Ann. N. Y. Acad. Sci. **83**:280–289.
21. **Gelber SE, Aguilar JL, Lewis KL, Ratner AJ.** 2008. Functional and phylogenetic characterization of vaginolysin, the human-specific cytolysin from *Gardnerella vaginalis*. J. Bacteriol. **190**:3896–3903.
22. **Germain M, Krohn MA, Hillier SL, Eschenbach DA.** 1994. Genital flora in pregnancy and its association with intrauterine growth retardation. J. Clin. Microbiol. **32**:2162–2168.
23. **Gilsdorf JR, Marrs CF, Foxman B.** 2004. *Haemophilus influenzae*: genetic variability and natural selection to identify virulence factors. Infect. Immun. **72**:2457–2461.
24. **Graham S, Howes C, Dunsmuir R, Sandoe J.** 2009. Vertebral osteomyelitis and discitis due to Gardnerella vaginalis. J. Med. Microbiol. **58**:1382–1384.
25. **Guindon S, et al.** 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. **59**:307–321.
26. **Haggerty CL, Hillier SL, Bass DC, Ness RB.** 2004. Bacterial vaginosis and anaerobic bacteria are associated with endometritis. Clin. Infect. Dis. **39**: 990–995.
27. **Hall BG, Ehrlich GD, Hu FZ.** 2010. Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. Microbiology **156**:1060–1068.
28. **Harper J, Davis G.** 1982. Cell wall analysis of *Gardnerella vaginalis*. Int. J. Syst. Bacteriol. **32**:48–50.
29. **Harwich MD, Jr, et al.** 2010. Drawing the line between commensal and pathogenic Gardnerella vaginalis through genome analysis and virulence studies. BMC Genomics **11**:375. doi:10.1186/1471-2164-11-375.
30. **Hidalgo-Grass C, et al.** 2006. A streptococcal protease that degrades CXC chemokines and impairs bacterial clearance from infected tissues. EMBO J. **25**:4628–4637.
31. **Hiller NL, et al.** 2010. Generation of genic diversity among *Streptococcus pneumoniae* strains via horizontal gene transfer during a chronic polyclonal pediatric infection. PLoS Pathog. **6**:e1001108. doi:10.1371/journal.ppat.1001108.
32. **Hiller NL, et al.** 2007. Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. J. Bacteriol. **189**:8186–8195.
33. **Hillier SL, et al.** 1995. Association between bacterial vaginosis and preterm delivery of a low-birth-weight infant. The Vaginal Infections and Prematurity Study Group. N. Engl. J. Med. **333**:1737–1742.
34. **Hogg JS, et al.** 2007. Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. Genome Biol. **8**:R103. doi:10.1186/gb-2007-8-6-r103.
35. **Hu FZ, Ehrlich GD.** 2008. Population-level virulence factors amongst pathogenic bacteria: relation to infection outcome. Future Microbiol. **3**:31–42.
36. **Huson DH, Bryant D.** 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. **23**:254–267.
37. **Jenkinson HF, Lamont RJ.** 2005. Oral microbial communities in sickness and in health. Trends Microbiol. **13**:589–595.
38. **Katoh K, Toh H.** 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinform. **9**:286–298.
39. **Klockgether J, Cramer N, Wiehlmann L, Davenport CF, Tümmler B.** 2011. *Pseudomonas aeruginosa* genomic structure and diversity. Front. Microbiol. **2**:150. 10.3389/fmicb.2011.00150.
40. **Konstantinidis KT, Tiedje JM.** 2005. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. U. S. A. **102**:2567–2572.
41. **Kosakovsky Pond SL, Frost SDW, Muse SV.** 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics **21**:676–679.
42. **Koumans EH, et al.** 2007. The prevalence of bacterial vaginosis in the United States, 2001-2004; associations with symptoms, sexual behaviors, and reproductive health. Sex. Transm. Dis. **34**:864–869.
43. **Lacross NC, Marrs CF, Patel M, Sandstedt SA, Gilsdorf JR.** 2008. High genetic diversity of nontypeable *Haemophilus influenzae* isolates from two children attending a day care center. J. Clin. Microbiol. **46**:3817–3821.
44. **Lagacé-Wiens PR, et al.** 2008. *Gardnerella vaginalis* bacteremia in a previously healthy man: case report and characterization of the isolate. J. Clin. Microbiol. **46**:804–806.
45. **Landan G, Graur D.** 2008. Local reliability measures from sets of co-optimal multiple sequence alignments. Pac. Symp. Biocomput. **2008**:15–24.
46. **Lander ES, Waterman MS.** 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics **2**:231–239.
47. **Lawrence JG.** 1997. Selfish operons and speciation by gene transfer. Trends Microbiol. **5**:355–359.
48. **Menard JP, Fenollar F, Henry M, Bretelle F, Raoult D.** 2008. Molecular quantification of *Gardnerella vaginalis* and *Atopobium vaginae* loads to predict bacterial vaginosis. Clin. Infect. Dis. **47**:33–43.
49. **Moncla BJ, Pryke KM.** 2009. Oleate lipase activity in *Gardnerella vaginalis* and reconsideration of existing biotype schemes. BMC Microbiol. **9**:78. doi:10.1186/1471-2180-9-78.
50. **Mukundan D, Ecevit Z, Patel M, Marrs CG, Gilsdorf JR.** 2007. Pharyngeal colonization dynamics of *Haemophilus influenzae* and *Haemophilus haemolyticus* in healthy adult carriers. J. Clin. Microbiol. **45**:3207–3217.
51. **Neri P, Salvolini S, Giovannini A, Meriotti C.** 2009. Retinal vasculitis associated with asymptomatic *Gardnerella vaginalis* infection: a new clinical entity. Ocul. Immunol. Inflamm. **17**:36–40.
52. **Ness RB, et al.** 2005. A cluster analysis of bacterial vaginosis-associated microflora and pelvic inflammatory disease. Am. J. Epidemiol. **162**:585–590.
53. **Nugent RP, Krohn MA, Hillier SL.** 1991. Reliability of diagnosing bacterial vaginosis is improved by a standardized method of Gram stain interpretation. J. Clin. Microbiol. **29**:297–301.
54. **Park PW, Rosenbloom J, Abrams WR, Rosenbloom J, Mecham RP.** 1996. Molecular cloning and expression of the gene for elastin-binding protein (ebpS) in *Staphylococcus aureus*. J. Biol. Chem. **271**:15803–15809.
55. **Pearson WR, Lipman DJ.** 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U. S. A. **85**:2444–2448.
56. **Piot P, et al.** 1984. Biotypes of *Gardnerella vaginalis*. J. Clin. Microbiol. **20**:677–679.
57. **Randis TM, Kulkarni R, Aguilar JL, Ratner AJ.** 2009. Antibody-based detection and inhibition of vaginolysin, the *Gardnerella vaginalis* cytolysin. PLoS One **4**:e5207. doi:10.1371/journal.pone.0005207.
58. **Redfield RJ, et al.** 2006. Evolution of competence and DNA uptake specificity in the Pasteurellaceae. BMC Evol. Biol. **6**:82. doi:10.1186/1471-2148-6-82.
59. **Rissman AI, et al.** 2009. Reordering contigs of draft genomes using Mauve Aligner. Bioinformatics **25**:2071–2073.
60. **Sá-Leão R, et al.** 2008. High rates of transmission and colonization by *Streptococcus pneumoniae* and *Haemophilus influenzae* within a day care center revealed in a longitudinal study. J. Clin. Microbiol. **46**:225–234.
61. **Sá-Leão R, et al.** 2006. Identification, prevalence and population structure of non-typable *Streptococcus pneumoniae* in carriage samples isolated from preschoolers attending day-care centres. Microbiology **152**(Pt. 2): 367–376.
62. **Shen K, et al.** 2005. Identification, distribution, and expression of novel genes in 10 clinical isolates of nontypeable *Haemophilus influenzae*. Infect. Immun. **73**:3479–3491.

63. **Sivadon-Tardy V, et al.** 2009. *Gardnerella vaginalis* acute hip arthritis in a renal transplant recipient. J. Clin. Microbiol. **47**:264–265.
64. **Snel B, Huynen MA, Dutilh BE.** 2005. Genome trees and the nature of genome evolution. Annu. Rev. Microbiol. **59**:191–209.
65. **Sobel JD.** 2005. What's new in bacterial vaginosis and trichomoniasis? Infect. Dis. Clin. North Am. **19**:387–406.
66. **Stajich JE, et al.** 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. **12**:1611–1618.
67. **Swidsinski A, et al.** 2008. An adherent *Gardnerella vaginalis* biofilm persists on the vaginal epithelium after standard therapy with oral metronidazole. Am. J. Obstet. Gynecol. **198**:97.e1-97.e6.
68. **Taha TE, et al.** 1999. HIV infection and disturbances of vaginal flora during pregnancy. J. Acquir. Immune. Defic. Syndr. Hum. Retrovirol. **20**:52–59.
69. **Taha TE, et al.** 1998. Bacterial vaginosis and disturbances of vaginal flora: association with increased acquisition of HIV. AIDS **12**:1699–1706.
70. **Tettelin H, et al.** 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. U. S. A. **102**:13950–13955.
71. **Tettelin H, Riley D, Cattuto C, Medini D.** 2008. Comparative genomics: the bacterial pan-genome. Curr. Opin. Microbiol. **11**:472–477.
72. **Totten PA, Amse R, Hale J, Holmes KK.** 1982. Selective differential human blood bilayer media for the isolation of Gardnerella (Haemophilus) vaginalis. J. Clin. Microbiol. **15**:141–147.
73. **Turroni F, et al.** 2010. Genome analysis of Bifidobacterium bifidum PRL2010 reveals metabolic pathways for host-derived glycan foraging. Proc. Natl. Acad. Sci. U. S. A. **107**:19514–19519.
74. **Udayalaxmi J, Bhat GK, Kotigadde S.** 2011. Biotypes and virulence factors of *Gardnerella vaginalis* isolated from cases of bacterial vaginosis. Indian J. Med. Microbiol. **29**:165–168.
75. **van Passel MWJ, Marri PR, Ochman H.** 2008. The emergence and fate of horizontally acquired genes in *Escherichia coli*. PLoS Comput. Biol. **4**:e1000059. doi:10.1371/journal.pcbi.1000059. doi:10.1371/journal.pcbi.1000059.
76. **Watts DH, Eschenbach DA, Kenny GE.** 1989. Early postpartum endometritis: the role of bacteria, genital mycoplasmas, and *Chlamydia trachomatis*. Obstet. Gynecol. **73**:52–60.
77. **Xie J, et al.** 2006. Identification of new genetic regions more prevalent in nontypeable *Haemophilus influenzae* otitis media strains than in throat strains. J. Clin. Microbiol. **44**:4316–4325.
78. **Yeoman CJ, et al.** 2010. Comparative genomics of *Gardnerella vaginalis* strains reveals substantial differences in metabolic and virulence potential. PLoS One **5**:e12411. doi:10.1371/journal.pone.0012411.